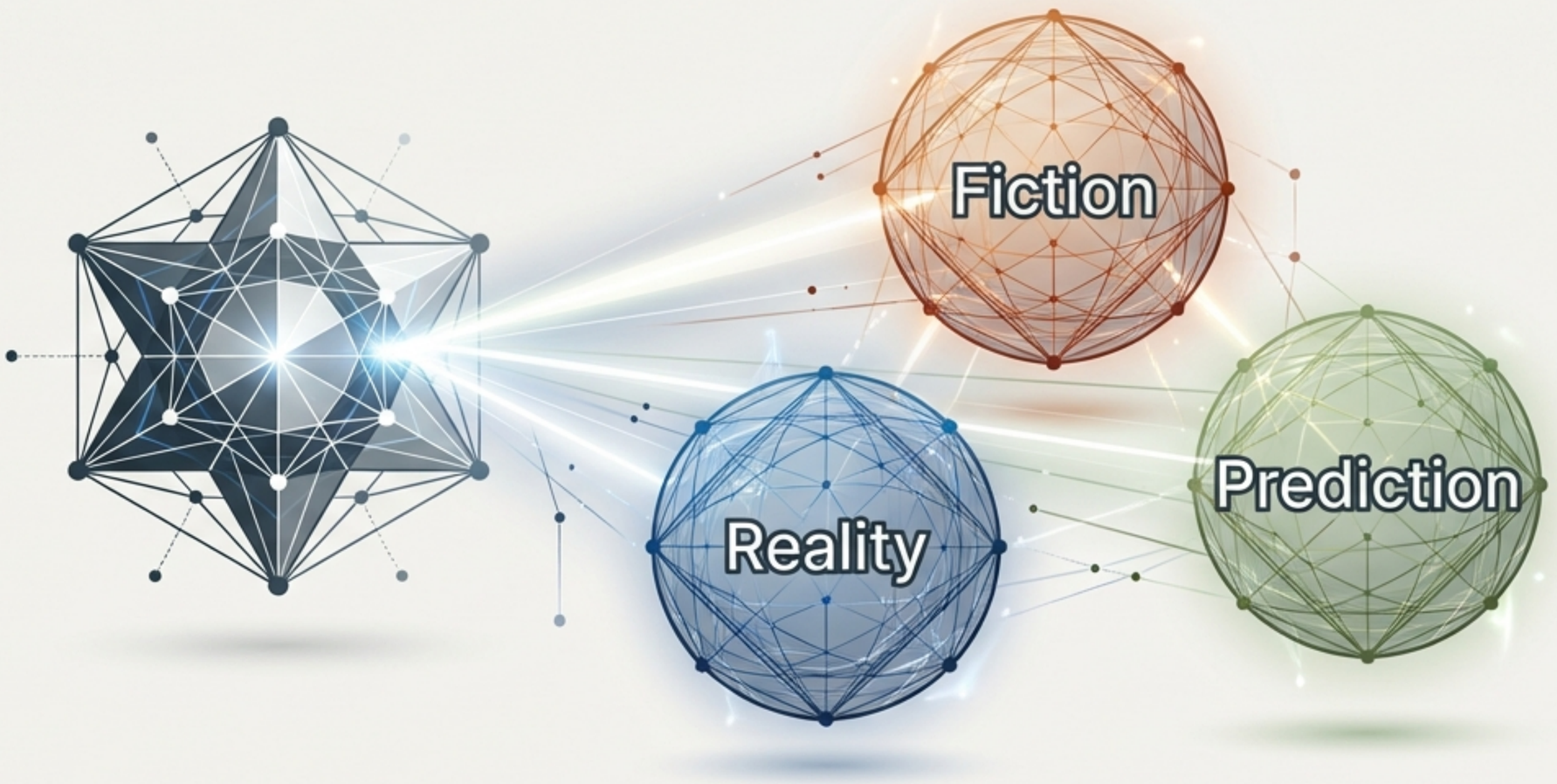
Making Sense of Hallucinations

Ontologies and Knowledge Graphs for Imaginable Reality



Based on the research by Vagan Terziyan, Svitlana Gryshko, Amit Shukla, Oleksandr Terziyan, and Oleksandra Vitko.

The Hallucination Paradox

In Artificial Intelligence

In AI, “hallucinations” are viewed as critical failures—bugs to be squashed. We penalize the system for generating non-veridical output.

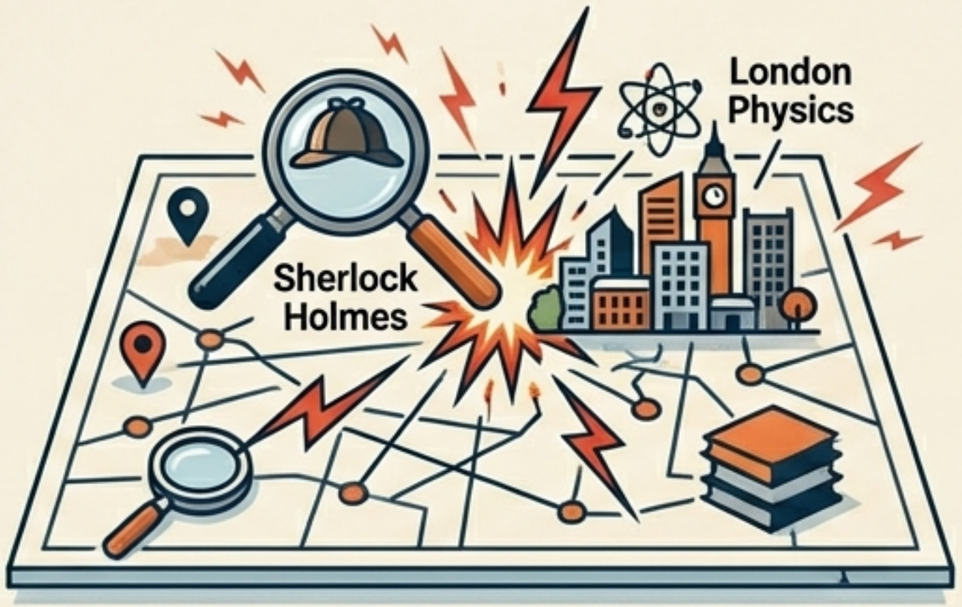
In Human Cognition

In humans, “offline world construction” is the engine of intelligence. The Default Mode Network (DMN) facilitates dreaming, future planning, and counterfactuals—all decoupled from sensory input.

Key Insight: The problem isn’t the imagination; it’s the lack of structural boundaries. We treat hallucinations as errors because our systems lack the architecture to manage them as features.

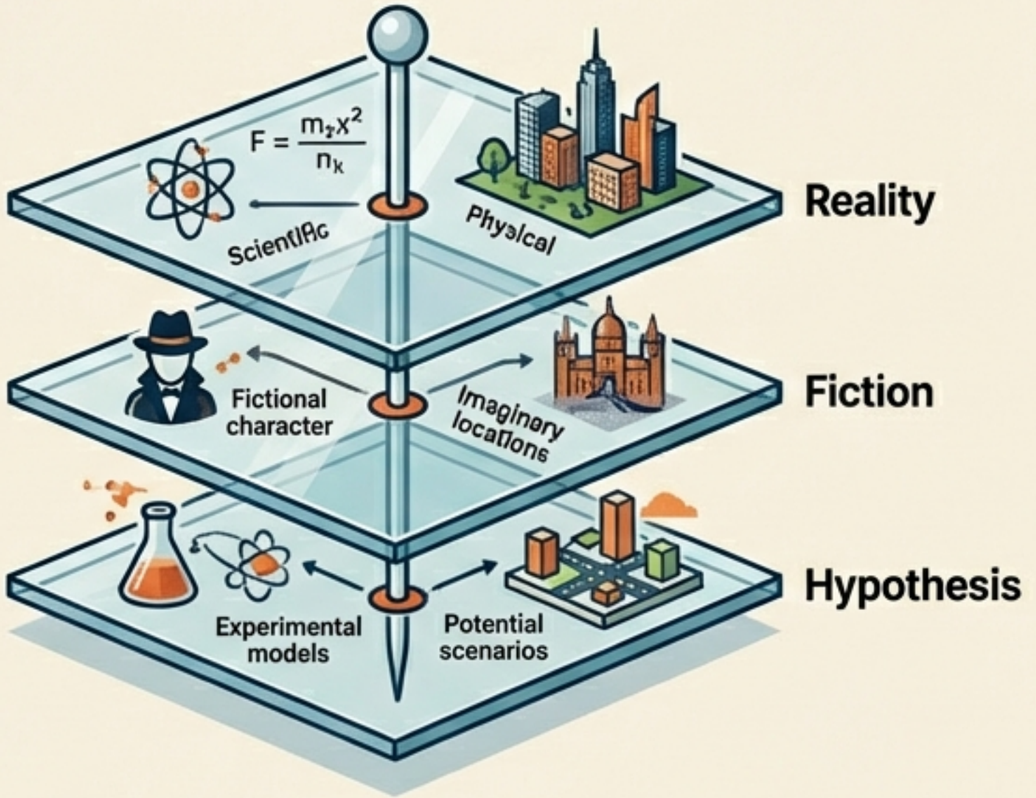
The ‘Single-World’ Trap of Current Semantics

Traditional Semantic Web (RDF/OWL)



Monolithic Truth: Speculation and Fact collapse into one flat layer, causing "Semantic Collapse".

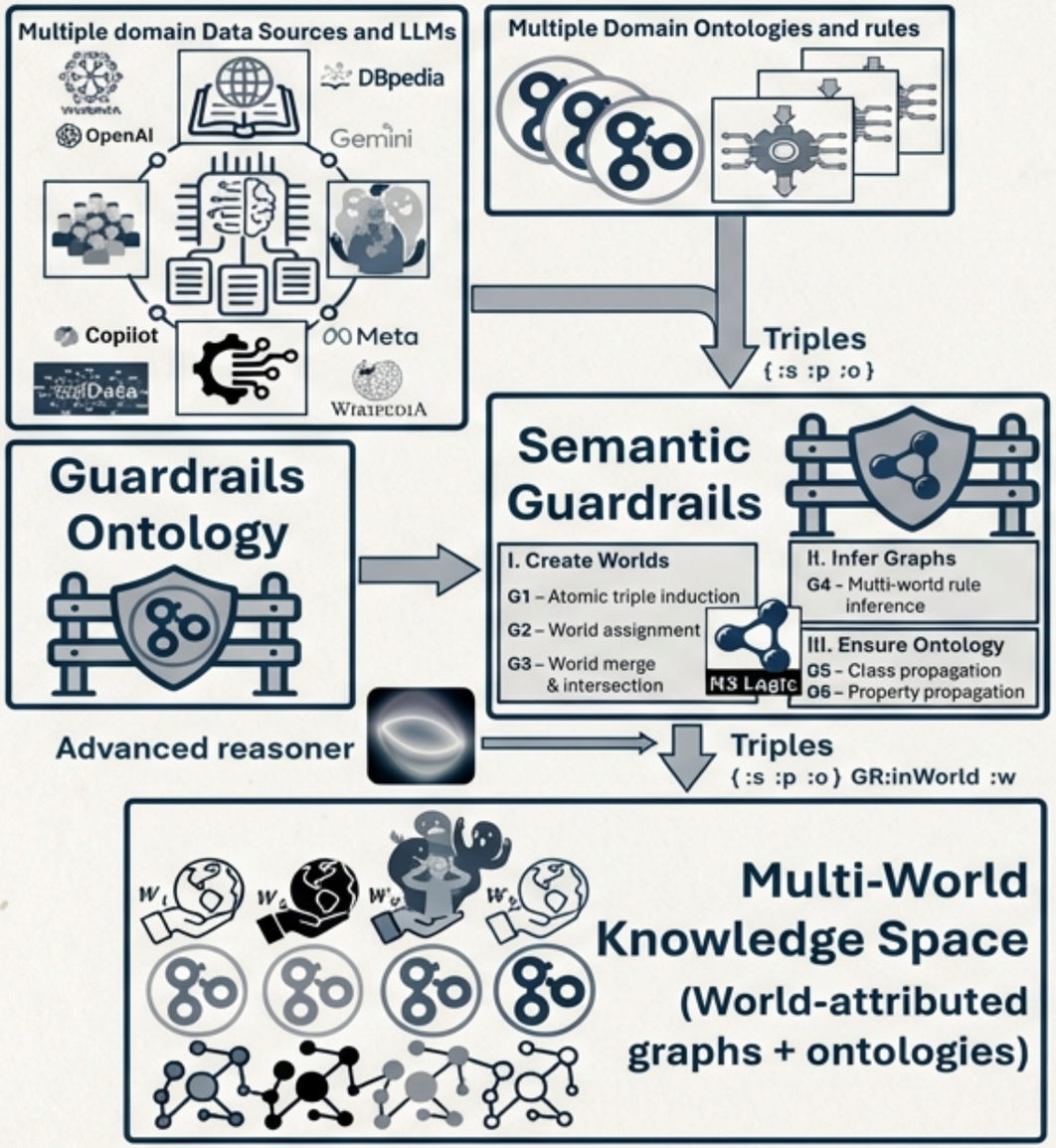
Ontological Pluralism



Multi-Reality: Distinct worlds coexist as first-class citizens. Contexts are ontologies, not just labels.

Introducing Semantic Guardrails

Epistemic Middleware for Hybrid Intelligence



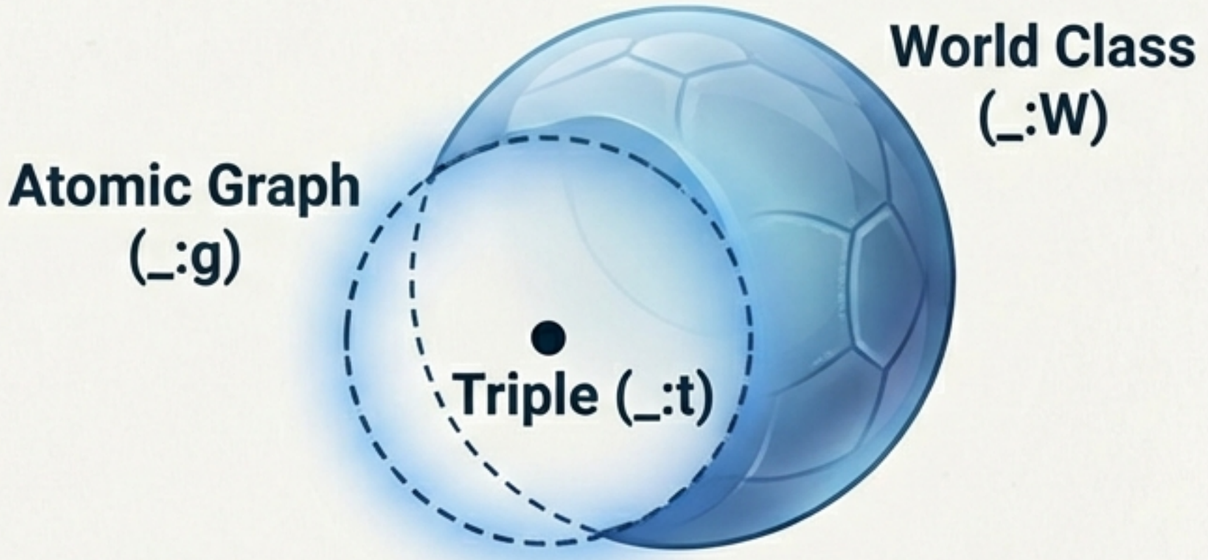
Definition: Semantic Guardrails are formal, ontology-level constraints that regulate inference across realities.

Core Functions:

1. **LIFT** triples into first-class graph objects.
2. **ORGANIZE** knowledge into world-relative structures.
3. **CONTROL** the propagation of inferences across worlds.

Step 1: Atomic Induction (Creating Worlds)

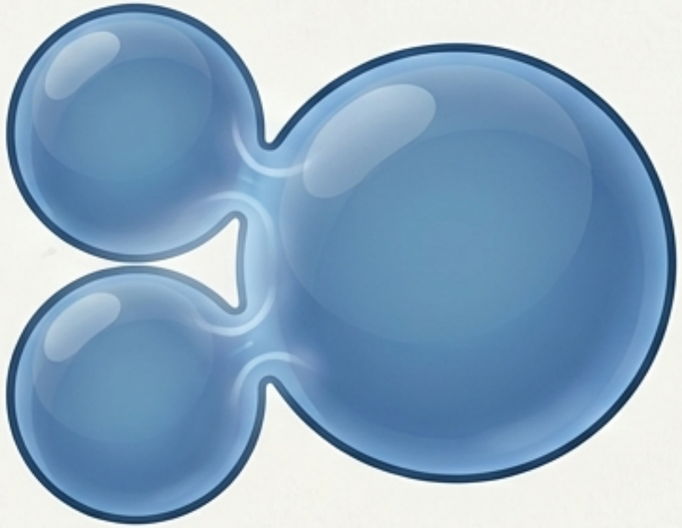
Triple (s, p, o) ⟶ Micro-World



The Mechanism (Guardrail 1):
We do not start with a global ‘base world.’ Every RDF triple creates a micro-world.
Input: A triple (e.g., ‘Holmes lives in London’).
Induction: This triple induces an Atomic Graph.
Emergence: This graph induces a new World Class.
Result: Worlds are dynamically synthesized from data, not predefined containers.

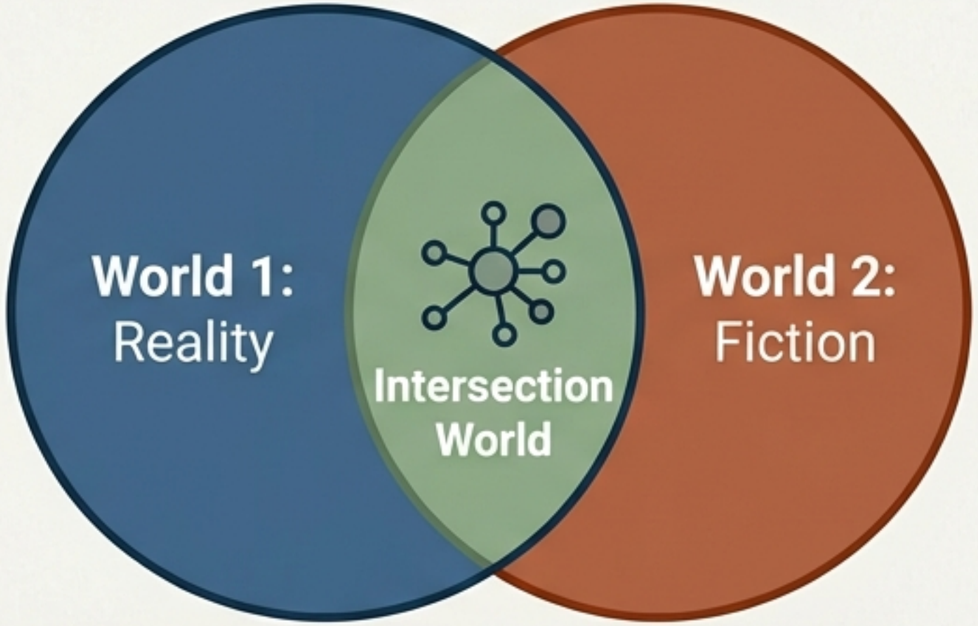
Step 2: Composition & Intersection

Mechanism A: Intra-World Merge



Condition: Graphs belong to the same world.
Action: Aggregation.

Mechanism B: Cross-World Merge

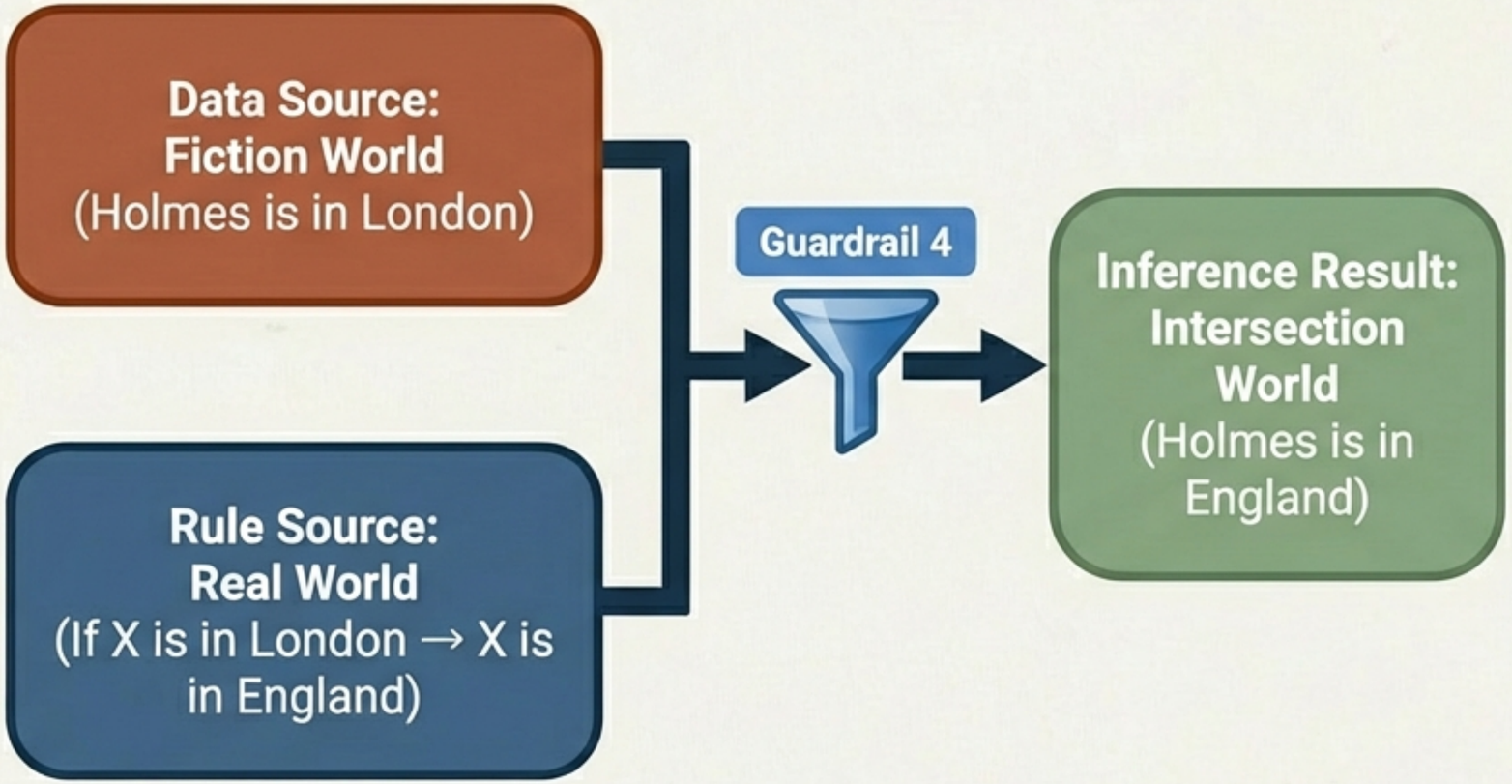


Condition: Graphs belong to different worlds.
Action: Create World Intersection ($W1 \cap W2$). The new composite graph lives *only* in this intersection.

"Guardrail 3 constructs a new reality that formally represents the overlap of incompatible assumptions."

Step 3: World-Preserving Inference

Logic flows through Guardrails, not around them.

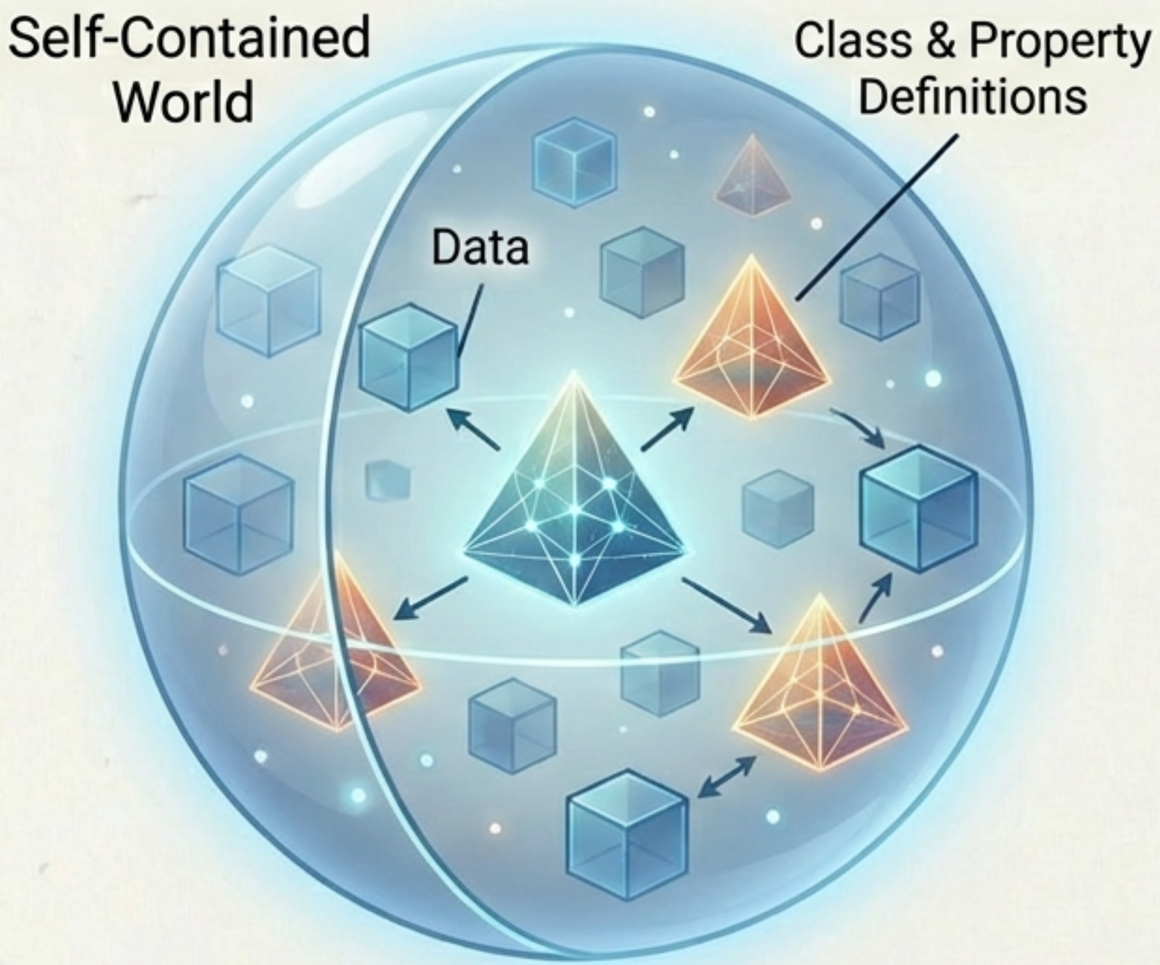


The Challenge: Applying real rules to fictional data without leaking fiction into reality.

The Solution:

- Rules have an origin world.
- When a Graph fires a Rule, Rule, the result goes to the Intersection.
- **Outcome**: Holmes is in England is **TRUE** only in the Reality ∩ Fiction intersection.

Step 4: Ontological Consistency



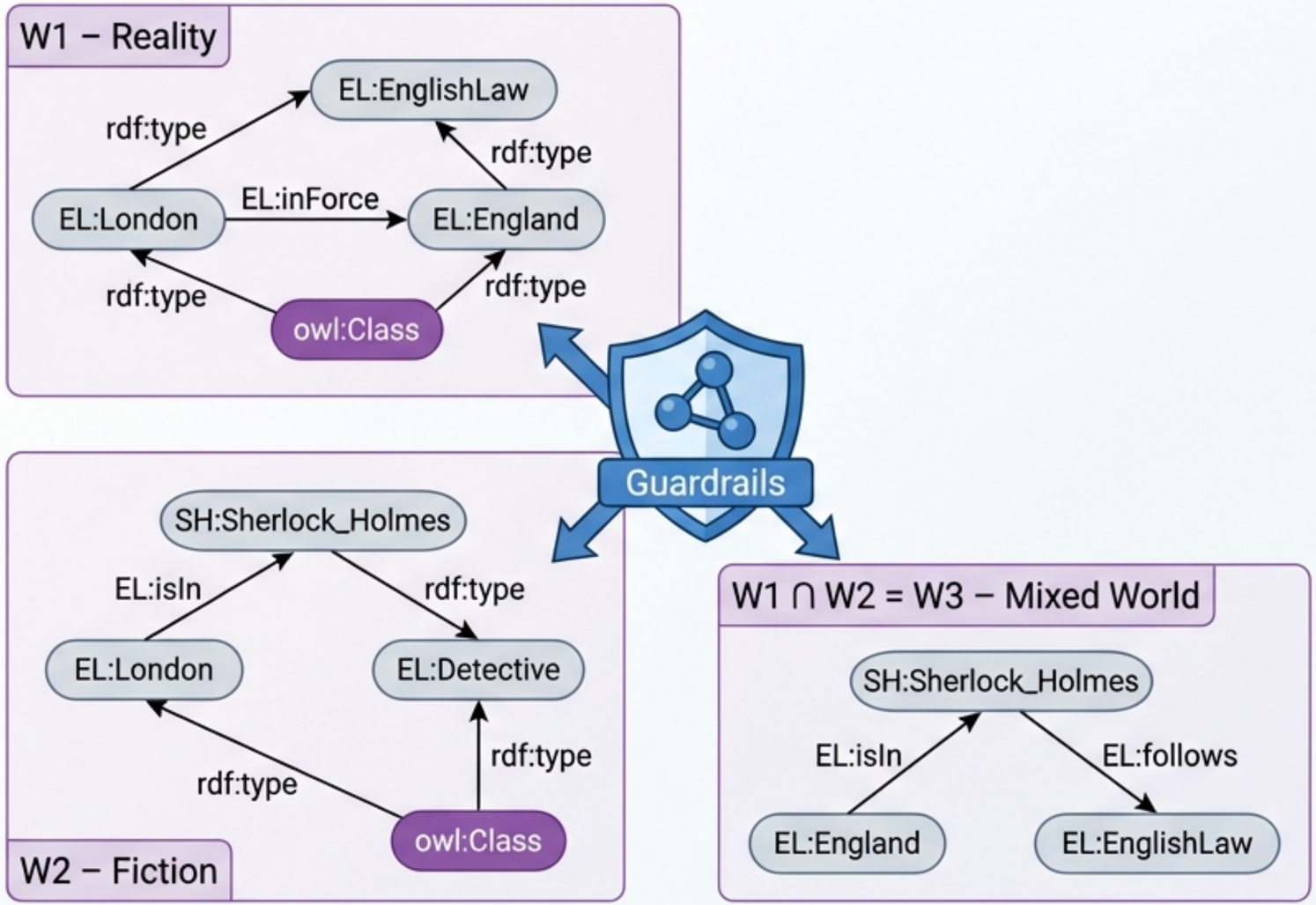
The Problem: If a fictional world uses a Class (e.g., 'Detective') defined in Reality, does it corrupt the real definition?

The Solution (Guardrails 5 & 6):

1. **Class Propagation**: Class declarations and subclass relations are lifted into the specific world.
2. **Property Propagation**: Domain, range, and inverse properties are world-scoped.

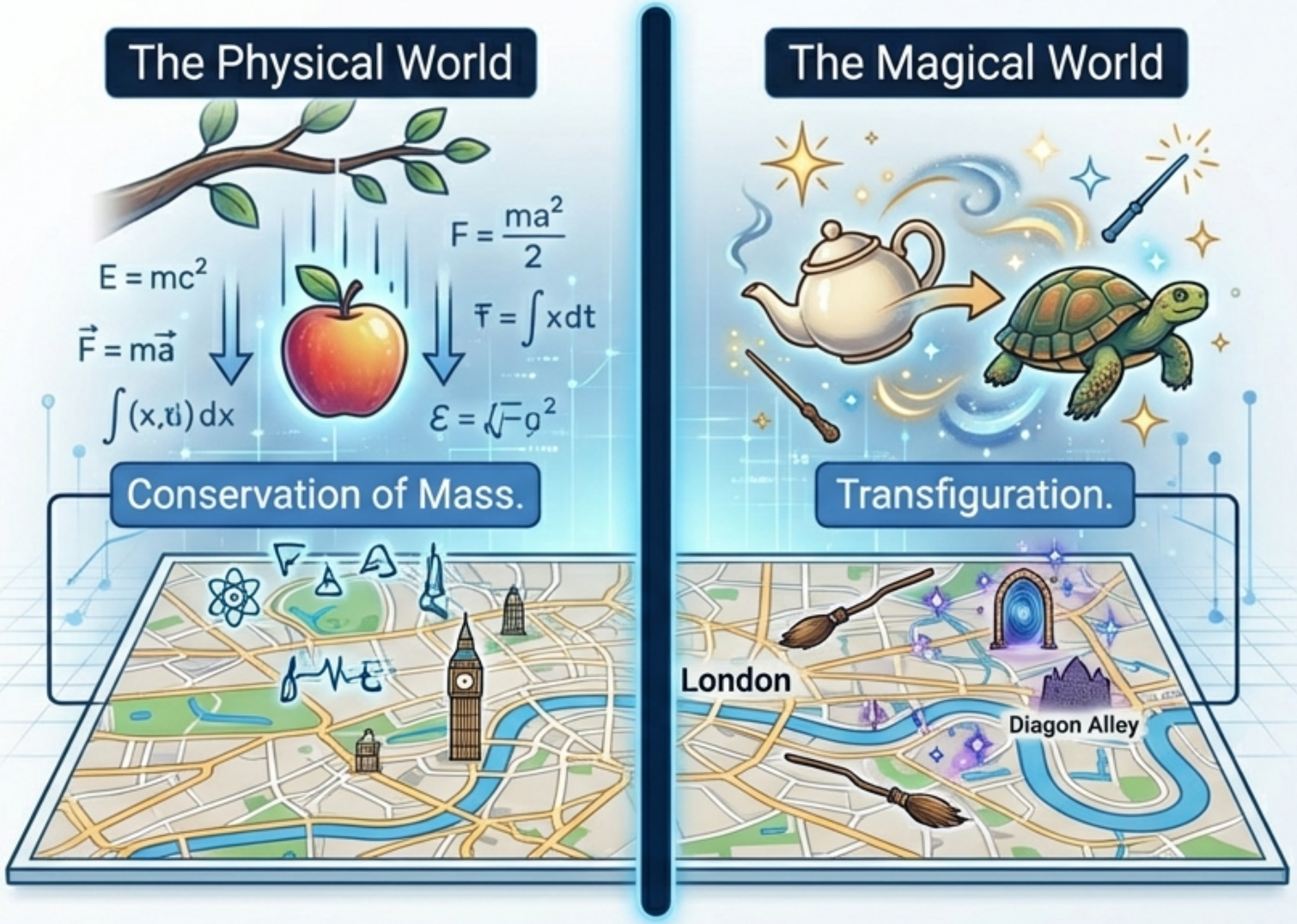
Result: A 'Law of Physics' in a magical world does not accidentally redefine physics in the actual world.

Proof Case I: Sherlock Holmes (Fiction ∩ Reality)



- **W1 (Reality):** Contains London, England, English Law.
- **W2 (Fiction):** Contains Sherlock Holmes, 'Holmes is in London'.
- **W3 (Intersection):** The Guardrails automatically generate this 'Mixed World'.
- **The Inference:** 'Holmes follows English Law' is derived here. It is not a real fact, nor purely fictional—it is a valid inference within the shared context.

Proof Case II: Harry Potter (Alternative Laws)



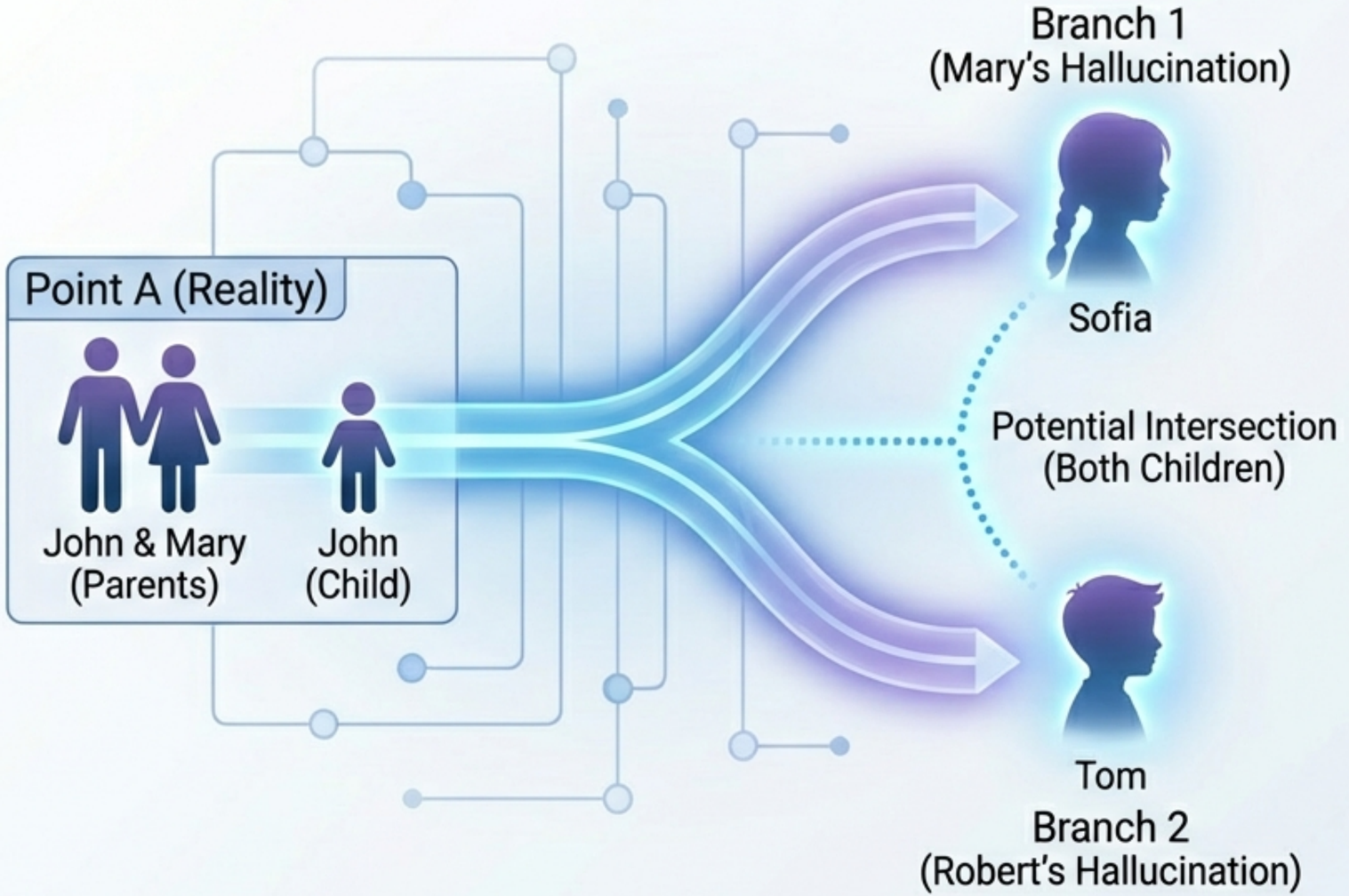
The Challenge: Handling incompatible ontologies (Magic vs. Physics).

The Guardrail Effect:

- **Physical** rules originate in **World_Physics**.
- **Magical** rules originate in **World_Magic**.
- The **Intersection** is restricted.

Result: Diagon Alley exists in London *only* within the Magical World. Magical spells do not corrupt physical databases.

Proof Case III: The Multiverse of Planning



The Multiverse of Planning

- **Scenario:** Parents imagining future children.
- **Mechanism:** These are 'Anticipatory Worlds.' They share a past but diverge in the future.
- **Outcome:** The system allows mutually exclusive futures to coexist. We can reason about 'Sofia' and 'Tom' simultaneously without them colliding or becoming 'fact' before they happen.
- **Application:** Robust Scenario Planning and Digital Twins.

Implementation: Logic as Code

Implementing Guardrails in N3 Logic

```
{
  ?R a GR:Rule ; GR:isOriginatedFromWorld ?w_rule .
  ?g a GR:Graph , ?w_graph ; GR:firesRule ?R .
}
=>
{
  _:W owl:intersectionOf ( ?w_rule ?w_graph ) .
  _:g_final a GR:Graph , _:W .
}
```

Logic Explanation: This snippet (Guardrail 4.4) demonstrates the 'Intersection' logic. It explicitly constructs a new world class (_:W) whenever a rule interacts with a graph from a different context. Source: Appendix II, Vagan Terziyan et al.

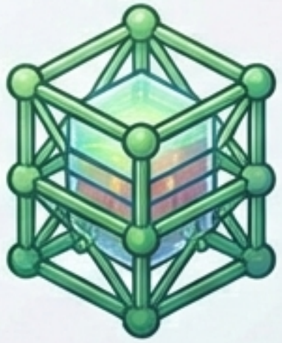
Implications for Generative AI

Guardrails as Epistemic Middleware

The Problem: LLMs lack world separation—they conflate fact, fiction, and style in one latent space.

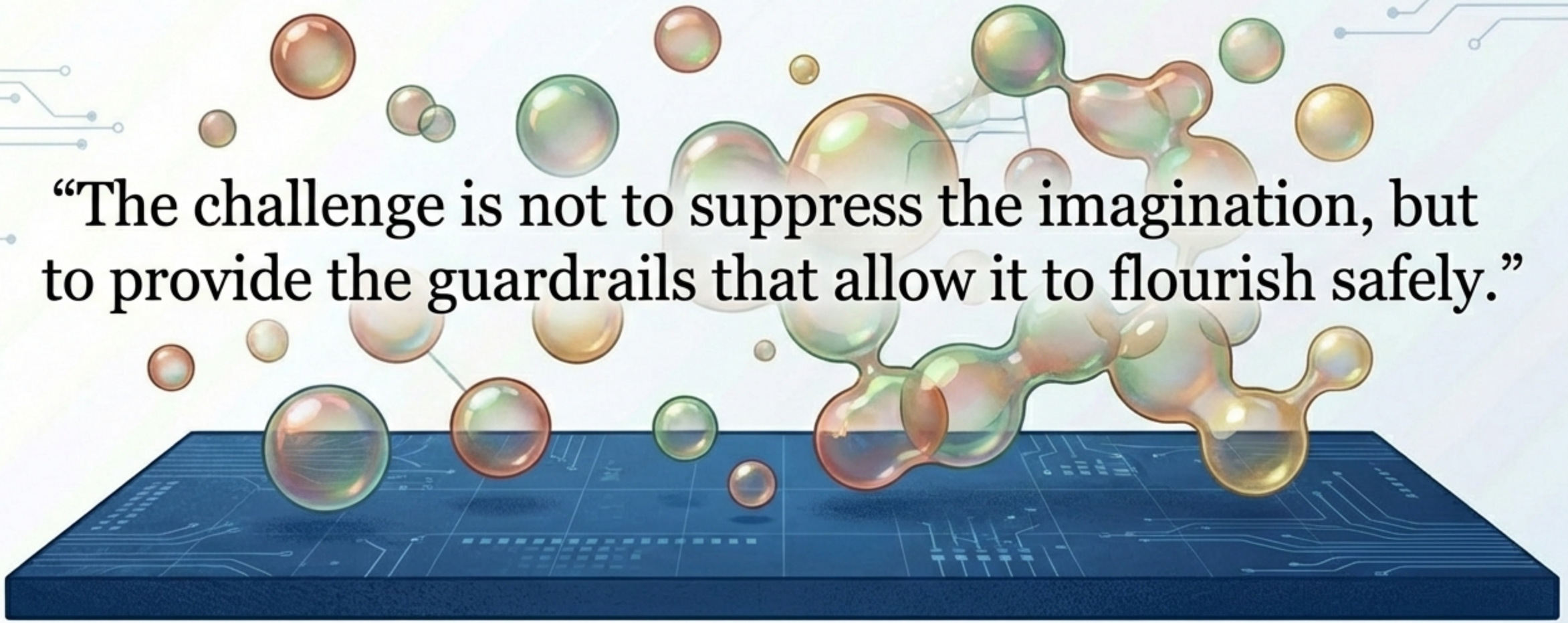
The Solution: Guardrails atomize output into triples and assign them to specific 'Hallucination Worlds.' Symbolic reasoning checks consistency within those bounded worlds.

From Suppression to Structure

Current Approach (Safety Filters)	Guardrails Approach (This Framework)
 • **Goal:** Eliminate hallucination. • **Method:** External database lookup / RAG. • **Result:** Constrained creativity, “I cannot answer that.”	• **Goal:** Managed imagination.  • **Method:** World-scoped graphs & Intersection logic. • **Result:** “Here is a valid inference **assuming* your hypothetical scenario.”

Key Concept: Hallucinations are treated as Manifestations of Unmanaged Imagination. Structure them, and they become intelligence.

The Horizon: World-Aware Intelligence



> “The challenge is not to suppress the imagination, but to provide the guardrails that allow it to flourish safely.”

Adopt multi-reality semantics to build safer, more cognitively aligned neuro-symbolic systems.